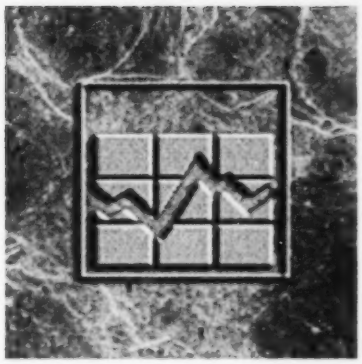
Catalogue no. 75F0002MIE — No. 010
ISSN: 1707-2840
ISBN: 0-662-41064-5

Research Paper

Income Research Paper Series

General Housing Imputation (excluding utilities) in the Survey of Labour and Income Dynamics (SLID)



by Georgina House

Income Statistics Division and Social Survey Methods Division
Jean Talon Building, Ottawa, K1A 0T6

Telephone: 613 951-7355

Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Income Statistics Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: (613) 951-7355; (888) 297-7355; income@statcan.ca).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 75F0002MIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Income Statistics Division

Income research paper series

General Housing Imputation (excluding utilities) in the Survey of Labour and Income Dynamics (SLID)

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2005

All rights reserved. The content of this publication may be reproduced, in whole or in part, and by any means, without further permission from Statistics Canada, subject to the following conditions: that it is done solely for the purposes of private study, research, criticism, review, newspaper summary, and/or for non-commercial purposes; and that Statistics Canada be fully acknowledged as follows: Source (or "Adapted from", if appropriate): Statistics Canada, name of product, catalogue, volume and issue numbers, reference period and page(s). Otherwise, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopy, for any purposes, without the prior written permission of Licensing Services, Marketing Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

July 2005

Catalogue no. 75F0002MIE, Vol. 10

Frequency: Occasional

ISSN: 1707-2840

ISBN: 0-662-41064-5

Ottawa

La version française de cette publication est disponible sur demande (nº 75F0002MIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

106-04671/

Table of Contents

Introduction 5
Longitudinal Imputation 6
Cross-sectional Donor Imputation 9
Dwelling type and tenure module 10
Owners and renters module 10
Owners module 11
Mortgage module 11
Property taxes and condominium charges module 11
Renters module 12
Rent module 12
Evaluation of the Imputation Methods for reference year 2002 13
Graph 1: Donor pool histogram for RENTM25 14
Graph 2: File after imputation (with RENTM25<1500) histogram for RENTM25 14
Conclusion 16
Appendix 1 18
Variable list 18

Introduction

For some time, Canada Mortgage and Housing Corporation (CMHC[1]) has used data on housing characteristics and housing-related expenditures from the Census of Population. Although the Census data source serves CMHC's purposes to a large extent, the federal government agency turned to the annual household surveys of Statistics Canada to provide information on a more frequent basis. This would allow them to have a better picture of annual trends, and perhaps have a greater choice of other characteristics with which to cross housing data on Canadian households. In 2001, CMHC began to sponsor additional content in both the Survey of Labour and Income Dynamics (SLID) and the Survey of Household Spending (SHS), starting with reference year 2002[2].

The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey initiated in 1993. The survey was designed to measure changes in the economic well-being of Canadians as well as the factors affecting these changes. The target population consists of all persons living in Canada with the following exclusions: persons living in Yukon, the Northwest Territories, and Nunavut, persons living on Reserves, persons living in institutions, and military personnel living in barracks.

The SLID sample is comprised of 2 panels. Each panel remains in the survey for 6 consecutive years and a new panel is rotated in every 3 years. In January following the reference year, SLID sample households are contacted by telephone interviewers. Demographic information is collected for every person in the household. Complete survey data are collected for every eligible person in the household over the age of 15. Questions are asked on labour (labour market activity, work experience, jobless spells, and job information), educational attainment and income sources. At the end of the January interview, respondents are informed that they will be contacted again in May when they will be asked to supply data on income as well as certain expense items. However, the respondent may elect to grant permission to Statistics Canada to retrieve all the data required from the T1 tax file, thereby avoiding the necessity of a second interview. Collection of income data is deferred until May so that the respondent will be more familiar with the required data (having just filed an income tax return).

Although originally designed as a longitudinal survey, SLID has always maintained the capability of producing cross-sectional estimates. All persons who are members of selected SLID households in the beginning of the first year of a panel's existence are longitudinal sample persons for SLID. As such, it is these individuals that are followed longitudinally. Any (non-longitudinal) person living in a household with a longitudinal person is referred to as a cohabitant. Cohabitants living with cross-sectionally eligible longitudinal persons will also be cross-sectional sample persons.

---

1. Canada Mortgage and Housing Corporation (CMHC) is a federal government agency with the mandate to promote: housing construction, repair and modernization; housing affordability and choice; improvements to overall living conditions; the availability of low-cost financing; and, the national well-being of the housing sector.
2. At the time of this report, this sponsorship is continuing every year.

For more information about survey concepts, definitions and design please refer to Statistics Canada publication: "*Survey of Labour and Income Dynamics - A survey overview*", http://www.statcan.ca:8096/bsolc/english/bsolc?catno=75F0011X

To fill CMHC's need for more data, over 20 housing related questions were added to the SLID labour interview. Some questions apply to homeowners and renters, some only to homeowners, and some only to renters. SLID started to collect information on mortgage payments, property taxes, condominium fees, rent payments and what is included in their rent.

Because of non-response to specific questions, imputation of housing related content was introduced in SLID[3]. The purpose of imputation is to replace erroneous or missing data with values that will provide reasonable estimates. Two methods of imputation were used, longitudinal imputation and cross-sectional donor imputation. They are discussed in the next two sections.

Longitudinal Imputation

Longitudinal imputation takes into account, as much as possible, information from the previous wave. The effectiveness of the imputation is greatly increased when data from consecutive waves are highly correlated.

Longitudinal imputation for the housing variables was performed wherever possible. Longitudinal imputation is preferred, especially when the household has not moved. We assume that if the household did not move residence from the previous year, then the information for this year would likely be the same as that collected in the previous year.

Since reference year 2002 was a new panel year, longitudinal imputation was performed differently depending on whether the household was part of the first wave of the new panel or not. For households not part of the first wave of the new panel, information for three housing variables (dwelling tenure, dwelling type and number of bedrooms) could be obtained from previous years. These were the only three housing variables that were not new to SLID in reference year 2002. (Mortgage on the dwelling was also available from past years, but it was decided not to do longitudinal imputation on this variable, because the mortgage status is more likely to change even if it appears that the address has not changed.) Missing information was imputed for households using information from that same household from previous years if the postal code remained the same. This was considered a good indication that the household had not moved.

The Labour Force Survey (LFS) and the Rent Survey, an LFS supplement, were used as sources of information for longitudinal imputation for those households part of the first

3. SLID has always done imputation of missing values for its income content.

wave of the new panel[4]. Missing information was imputed for households using information from that same household from the LFS data if the postal code remained the same. To use the Rent Survey, it was necessary to ensure that the dwelling tenure variable on SLID indicated that the household was still a renter.

Longitudinal imputation was successful in lowering the number of households with missing values for variables which were available from previous SLID years. For others, however, it had little to no impact. The new housing variables for which the only source of information was the rent survey had very few successful longitudinal imputations. The main reason for this was that the Rent Survey was only available for households in panel four who were considered renters. Of this small number, only those with valid data on the Rent Survey and whose postal codes matched were then imputed.

For reference year 2003 longitudinal imputation was done on all housing variables for both panels. Missing information was imputed for households using information from that same household from reference year 2002 if the postal code remained the same. If dwelling tenure was available for reference year 2003 then missing information was only imputed if the tenure also matched.

Table 1 contains the weighted percentage of missing values for each variable. Where longitudinal imputation was possible the weighted percentage of missing values after longitudinal imputation is also included. Since longitudinal imputation was only available for certain variables in 2002 some values are missing. Some variables apply only to certain households (i.e. homeowners and renters, only homeowners, and only renters); as a result many have different denominators. Those who require a value other than 'not applicable' to be imputed make up these different denominators. For variable definitions see appendix 1.

4. SLID uses households rotating out of the LFS in December SLID reference year minus one and January SLID reference year.

Table 1- Impact of longitudinal imputation for 2002 and 2003

Variable	Group of households that require imputation for this variable	Missing Values (%)				Number of households longitudinally imputed	
		Before		After			
		2002	2003	2002	2003	2002	2003
dwldet25	Owners and Renters	6.9	7.4	0.6	0.5	1646	1871
dwtenr25	Owners and Renters	6.6	7.4	0.4	0.5	1637	1908
dwltyp25	Owners and Renters	6.9	7.4	0.6	0.5	1646	1871
rooms25	Owners and Renters	6.9	8.0	2.7	0.6	1215	2031
opbu25	Owners and Renters	8.1	9.4	8.0	0.8	9	2649
rnre25	Renters	8.8	10.8	8.7	0.9	7	798
rnpk25	Rent payers	8.0	9.5	7.9	0.6	6	689
rnht25	Rent payers	8.0	9.5	7.9	0.6	4	689
rnwa25	Rent payers	8.0	9.5	7.9	0.6	5	689
rnec25	Rent payers	8.0	9.5	7.9	0.6	2	689
rntv25	Rent payers	8.0	9.5	7.9	0.6	1	689
rnfg25	Rent payers	8.0	9.5	7.9	0.6	5	689
rnst25	Rent payers	8.0	9.5	7.9	0.6	6	689
rnwd25	Rent payers	8.0	9.5	7.9	0.6	3	689
rnfu25	Rent payers	8.0	9.5	8.0	0.6	0	689
rnno25	Rent payers	8.0	9.5	8.0	0.6	0	689
rentm25	Renters	12.1	14.4	10.2	2.6	109	946
repa25	Owners and Renters	7.0	7.8	n/a[2]	0.6	n/a	1987
heat25	Owners and Renters	11.6	12.8	n/a	2.3	n/a	2703
heatg25	Owners and Renters	11.6	12.8	n/a	2.3	n/a	2703
opfm25	Owners and Renters	3.4[1]	7.0	n/a	0.7	n/a	1716
mortg25	Owners	7.7	8.1	n/a	0.8	n/a	1426
cond25	Owners	6.5	7.1	n/a	0.6	n/a	1267
mortgn25	Mortgage payers	9.5	10.3	n/a	1.9	n/a	861
mortgm25	Mortgage payers	29.4	29.5	n/a	8.6	n/a	2297
prtxm25	Owners	25.7	26.7	n/a	5.4	n/a	4546
condm25	Condominium members	13.8	15.7	n/a	5.4	n/a	95
rnbs25	Renters	9.6	11.3	n/a	1.2	n/a	834

1. In reference year 2002 only households (outside of New Brunswick) who have a postal code indicating rural (second digit of postal code= 0) were asked if the dwelling operated a farm on the property (OPFM25), all others are set to 'not applicable'. The true non-response rate is 16.6% with a denominator of 11,018.

2. Longitudinal imputation not available in 2002.

Cross-sectional Donor Imputation

Donor imputation consists of identifying a group of households sharing several characteristics with the household to be imputed, then selecting one of them to be the donor. The value reported by the donor replaces the missing field of the imputed household. For example, SLID imputed the income of an individual by picking the income of a person living in the same province and with the same age range, sex, level of education, type of job (employee, self-employed) and occupation.

Donor imputation was performed for the remaining variables and for all those households where longitudinal imputation was not available. Donor imputation was broken down into seven modules. This was done in order to facilitate the use of imputed variables from one module as auxiliary or matching variables in another. Imputed records may also act as donors in successive rounds of imputation.

Each module has different lists of auxiliary and matching variables. Auxiliary variables are used to create imputation groups[5], whereas matching variables are used in a score function to find the best suited donor within an imputation group. In some cases, auxiliary variables created receiver groups[6] with no matching donor groups[7]; when this was the case, collapsing of auxiliary variables was performed. For example, for all but one module, the province variable was collapsed into five regions. When collapsing was not sufficient enough to create possible donors for all receivers, auxiliary variables were then turned into matching variables. Within each module, variables on the "variables to be imputed" list were also used as matching variables when a value for that variable was available. For example, in the Dwelling type and tenure module, if dwelling type was missing and dwelling tenure was not, then it was attempted to find a donor with a matching value of dwelling tenure. Within each imputation group, donors were evaluated based on a score function. This score function was set up based on the score function used in SLID income imputation:

$$s(X,Y) = \sum_{k=1}^{K} p_k \, \mathrm{I}(X_k, Y_k), \qquad \text{where } \mathrm{I}(X_k, Y_k) = \begin{cases} 1 & \text{if } X_k = Y_k \\ 0 & \text{if not.} \end{cases}$$

Note that p_k is a weight allowing us to assign more or less importance to the matching variable k. It was decided that all matching variables are of equal importance, hence $p_k=1$. X_k is the value of the receiver's variable k and Y_k is the value of the donor's variable k.

5. Both donors and receivers (records requiring imputation) are separated into imputation groups. Every receiver within an imputation group is compared to every donor within the same imputation group.
6. Receiver groups are the groups, based on auxiliary variables, in each module of records requiring imputation.
7. Donor groups are the groups, based on auxiliary variables, in each module of possible donor records.

A donor is then selected based on this score. The donor with the highest score is selected. If there are several donors with the high score value then a donor is randomly chosen out of this list. For those modules where there are no matching variables a donor is selected randomly from the entire imputation group.

The following is a breakdown of each module. Each table contains a list of variables to be imputed along with lists of auxiliary and matching variables. It was decided that for modules in which imputation of numeric variables was required, those that fell in the top one percent would be excluded from the donor pool. This occurs in four different modules; the Owners and renters module, the Mortgage module, the Property tax and condominium charges module and finally the Rent module. One reason for this was to lessen the chance of imputing a value higher than household income for dollar amounts such as rent or mortgage.

Dwelling type and tenure module

The Dwelling type and tenure module is the first module to run. This is done because every proceeding module uses either dwelling type or dwelling tenure as an auxiliary or matching variable. Note that the number of codes is in brackets.

Records entering module:	All households
Variables to be imputed:	Detailed dwelling type, Type of dwelling, Dwelling tenure
Auxiliary variables:	Province (10), Urban size group (3), Number of persons in the household as at December 31 (3)

Owners and renters module

The Owners and renters module is the next to run. Variables to be imputed in this module are variables for which all records require a value other than 'not applicable'. When choosing possible donors in this module, all those who had a value of 9 or higher for the number of bedrooms were excluded.

Records entering module:	All households
Variables to be imputed:	Number of bedrooms, Dwelling repairs needed, Principal heating fuel, Principal heating fuel group, Household operates farm, Household operates business
Auxiliary variables:	Province (5), Urban size group (3), Type of dwelling group (3), Number of persons in the household as at December 31 (3)
Matching variables:	Ownership of dwelling (2)

Owners module

The next module to be run is the Owners module. This module uses imputed variables from both previous modules as auxiliary variables. All households who have a value of dwelling tenure indicating ownership of the dwelling enter into this module. All other records with missing values are set to 'not applicable'.

Records entering module:	Owners
Variables to be imputed:	Mortgage on the dwelling, More than one mortgage on the dwelling, Dwelling is part of condominium development
Auxiliary variables:	Province (5), Urban size group (3), Type of dwelling group (2), Number of bedrooms (4)
Matching variables:	Number of persons at December 31 (3)

Mortgage module

The Mortgage module follows the Owners module. This is done to ensure that only households who have a mortgage on the dwelling (MORTG25) enter into this module. All others with missing values are set to 'not applicable'. When choosing possible donors in this module, all those who had a value of monthly amount of regular mortgage payment higher than $3,000 were excluded.

Records entering module:	Owners with a mortgage on the dwelling
Variables to be imputed:	Monthly amount of regular mortgage payments
Auxiliary variables:	Province (5), Type of dwelling (2), Number of bedrooms (4), More than one mortgage on the dwelling (2)

Property taxes and condominium charges module

The Property taxes and condominium charges module can be run in succession with the Mortgage module. All homeowners enter into this module. All others with missing values are set to 'not applicable'. Since the variable, dwelling is part of a condominium complex (COND25), is one of the matching variables, homeowners who are not a part of a condominium complex will have their monthly condominium charges imputed to 'not applicable'. When choosing possible donors in this module, all those who had a monthly amount of property taxes higher than $500 or a value of monthly condominium charges higher than $800 were excluded.

Records entering module:	Owner
Variables to be imputed:	Monthly amount of property taxes, Monthly condominium charges

Auxiliary variables:	Province (5), Type of dwelling group-owners (2), Dwelling is part of a condominium development (2)
Matching variables:	Urban size group (3), Number of bedrooms (4)

Renters module

The Renters module must follow the Owners and renters module. All households who have a value of dwelling tenure indicating that they are a renter enter into this module. All others with missing values are set to 'not applicable'. Those who are renters, but do not pay a monthly rent (RENTM25=0) have variables such as rent includes electricity, parking, etc. are set to 'not applicable'.

Records entering module:	Renters
Variables to be imputed:	… included in rent (several items), Rent includes fully or partially furnished, Reduced rent reason, Rent calculated on the basis of income
Auxiliary variables:	Province (5), Urban size group (3), Type of dwelling group-renters (3)
Matching variables:	Number of bedrooms (4)

Rent module

Finally, the last module is the Rent module. All households who pay a monthly rent enter into this module. All other households with missing values are set to 'not applicable'. When choosing possible donors in this module, all those who had a monthly amount of rent higher than $1,500 were excluded.

Records entering module:	Renters
Variables to be imputed:	Monthly amount of rent
Auxiliary variables:	Province (5), Type of dwelling group–renters (3), Number of bedrooms (4), Reduced rent reason(2)
Matching variables:	…included in rent (several items), Number of bedrooms (4)

There was a wide range of percentages of missing values prior to donor imputation for reference year 2002. Certain variables have a relatively high percentage of missing values resulting in small donor pools. Table 1 contains the weighted percentage of missing values before cross-sectional imputation for both reference years 2002 and 2003.

Evaluation of the Imputation Methods for reference year 2002

To evaluate the success of the imputation methods, before and after frequencies for categorical variables and before and after means for continuous variables were examined. Changes of less than 1% in before and after percentages and a less than 6% change in before and after means were considered successful. A cross validation analysis was also performed by using 1,000 records with complete data. These records were imputed. Then, the original data was compared with the imputed data.

After imputation was completed using the methods described above, several analyses were done. These analyses were performed in order to ascertain whether or not the resulting data was similar to the original data. Before and after frequencies were taken on all imputed categorical variables. As an example, table 2 contains the before and after percentages for dwelling type. Prior to donor imputation, 4% of the households were missing this variable. Table 2 shows that the before and after percentages are quite similar. Any difference that does exist is less than 1% which is very reasonable.

Table 2: Percentage before and after donor imputation for dwelling type

DWLDET25 Code set	Before donor imputation (excluding missing values in the denominator)	After donor imputation
01 Single detached	66.5	65.9
02 Double	3.53	3.50
03 Row or terrace	4.13	4.14
04 Duplex	4.02	4.02
05 Low-rise apartment	14.0	14.4
06 High-rise apartment	5.01	5.17
07 Hotel, rooming house, camp	0.26	0.27
08 Mobile Home	2.42	2.42
09 Other	0.15	0.17

It is important to also take a close look at variables for which the percentage of missing values is much higher. As an example, table 3 contains the before and after percentages for principle heating source. Prior to donor imputation, 24% of the households were missing this variable. Once again, however, table 3 shows that the difference in percentages is always less than one percent. These results are typical of what was seen with all categorical variables

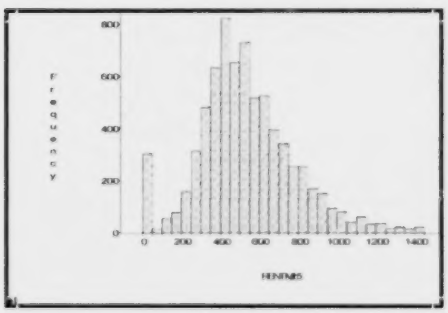
Table 3: Percentage before and after donor imputation for principle heating source

HEAT25 Code set	Percentage before donor imputation (excluding missing values in the denominator)	Percentage after donor imputation
01 Oil or other liquid fuel	17.1	16.7
02 Piped gas (natural gas)	41.0	41.8
03 Bottled gas (propane)	0.90	0.88
04 Electricity	33.0	33.5
05 Wood	7.39	7.07
06 Other	0.06	0.06

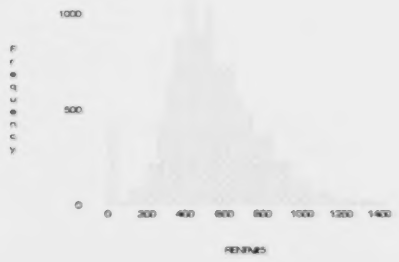
When looking at continuous variables, the before and after means were studied. This gave a good indication as to whether or not the data has been changed significantly. As an example of this, look at the before and after means of amount of monthly rent (RENTM25) and property taxes (PRTXM25). The mean prior to imputation of RENTM25 was $604.12 with 29% of households missing this variable; after imputation, the mean has dropped slightly to $589.53. This is a 2% reduction. The mean prior to imputation of PRTXM25 was $160.68 with 35% of households missing this variable; after imputation, the mean has again dropped slightly to $154.53. This is a 4% reduction. These results are characteristic of all continuous variables. The reductions range from 2 to 6 percent. A drop in the mean is to be expected, largely due to the fact that the top one percent of the donor pool was excluded.

A visual check was also performed. Continuous variables were plotted in a histogram to compare the donor pool to the newly imputed file. In order to make the comparison simpler, the top 1% of the data from the histogram of the newly imputed file was excluded. The histograms in graphs 1 and 2 show that the distribution before and after imputation is very similar. Histograms for all continuous variables were compared in this way and the example chosen is typical of what was seen. This is a good indication that the imputation methods chosen were successful in that they did not alter the data to any great extent.

Graph 1: Donor pool histogram for RENTM25



Graph 2: File after imputation (with RENTM25<1500) histogram for RENTM25



As an added analysis, a cross validation was done. A cross validation is a comparison on households for which we have complete data between values imputed using the methods described above and reported values. A cross validation is useful because it simulates the imputation process in a way that makes it easy to test its accuracy. A set of 1,000 records for which complete data was available was imputed using the methods described above. The original data for these records was compared with the newly imputed data. This cross validation comparison was only done on categorical variables. The data was imputed with an incorrect value between 10 and 40 percent of the time. This cross validation was run several times and the results below are characteristic of what was seen.

When looking at the results of the cross validation, one must keep in mind the fact that the imputation is done in sequential modules. The Dwelling type and tenure module is run first, thus the 11% of households with incorrect dwelling tenure will then be incorrect for most of the other modules. For example, those who are considered owners are the only ones that would be imputed a value for mortgage on the dwelling (MORTG25) all others would be given a value of not applicable.

For comparison purposes, random imputations were performed on the same 1,000 households. Values were randomly assigned to each variable based on the initial frequencies.

Since imputation is done sequentially, the imputation of one module is done before the imputation of another module. This generates two ways to proceed with the analysis. The imputation could have used the values created during the cross validation from the previous file, or could have used the new values that were randomly assigned from the previous file. In order to maintain a logical analysis of the variables being imputed, we use the cross validation data. In doing so respondents who were originally renters stayed renters, which meant that variables applicable to renters, like rent includes heat (RNHT25) would have a valid imputed value rather than a 'not applicable' which would happen if their randomly imputed value in a previous module changed them to an owner. Therefore those who are considered incorrectly imputed by the random method are based on the same subset of the population as those in the cross validation.

From these results, it can be seen that there is a difference between variables for which longitudinal imputation was successful and variables for which it was not. For example, random imputation was performed on DWTENR25. This was done several times and typically the data was imputed incorrectly 40% of the time, which is much higher than when the imputation methods discussed in this paper are used. As an example of a random imputation on a variable where longitudinal imputation was not available, look at the variable MORTG25. The results show that the imputation method produces only a slight improvement over the random imputation method. Typically, the incorrect value was imputed 43% of the time. It would appear from these results that the longitudinal imputation method is much better than random imputation, but that the cross-sectional donor imputation method is only slightly better. Table 4 is a table of the percentage

difference between the original data and the imputed data for the cross validation along with typical results of the random imputations.

Table 4: Percentage of incorrectly imputed categories for two imputation strategies

Variable	Cross sectional imputation (%)	Random imputation (%)
Dwtenr25	11	40
Dwldet25	20	54
Mortg25	40	43
Repa25	35	36
Heat25	39	67
Heatg25	39	67
Obpu25	16	16
Mortgn25	38	38
Cond25	13	17
Rnpk25	17	18
Rnwa25	16	19
Rnec25	19	21
Rnht25	20	22
Rntv25	15	15
Rnfg25	16	20
Rnst25	16	21
Rnwd25	16	17
Rnfu25	13	13
Rnno25	12	12
Rnre25	17	18
Rnbs25	17	19

Conclusion

Based on the initially high item non-response for housing variables, it was useful to put in place an imputation methodology. One of the strengths of the described imputation procedure is that the end result is a dataset with no missing values and internally consistent imputed records.

From these analyses, one can see that the imputation methods chosen were quite successful. Before and after frequencies for categorical variables and before and after means for continuous variables show only relatively small changes less than 1% and 6% respectively. As an added confirmation, before and after histograms are very similar in appearance. This was a confirmation that the imputation procedure did a good job in preserving univariate distributions.

A more challenging question is to what degree multivariate distributions were altered. To address this question, we looked at estimates for specific groups. For example, since age is not explicitly taken into account in the imputation, an existing correlation between age

and dwelling tenure may be lost. Analyses performed on the data in reference to the age of the oldest household member indicate that the imputation process does not affect the estimates of home owners based on age. The distribution of the age of the oldest member of the household of home owners is very similar before and after imputation. The average age of the person with the highest net income in households with a monthly mortgage for imputed versus non-imputed households is also quite similar. This is yet more evidence to suggest that the imputation method is quite good. Since it is impossible to test all variables for which there may be a correlation, one can't be sure that all estimates for specific groups are not affected by what appears to be, according to the cross validation analysis, a high error rate.

The bottom line is that, although the measurement error rate for some variables is high due to weakness in the underlying imputation models, the impact on univariate and some multivariate distributions seems to be quite limited. In other words, although initial non-response rates were high, the missing values were replaced without causing great changes to the overall estimates.

Appendix 1

Variable list

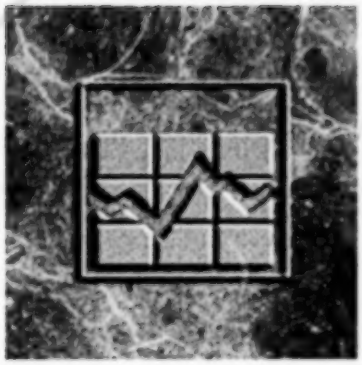
Variable	Variable Description	Module	Group of households that require imputation for this variable
dwldet25	Detailed type of dwelling	Dwelling type/Tenure	Owners and renters
dwtenr25	Ownership of dwelling	Dwelling type/Tenure	Owners and renters
dwltyp25	Dwelling type	Dwelling type/Tenure	Owners and renters
rooms25	Number of bedrooms	Owners/renters	Owners and renters
opbu25	Business on this property	Owners/renters	Owners and renters
repa25	Repairs needed	Owners/renters	Owners and renters
heat25	Principal heating fuel	Owners/renters	Owners and renters
heatg25	Principal heating fuel group	Owners/renters	Owners and renters
opfm25	Farm on this property	Owners/renters	Owners and renters
mortg25	Mortgage on the dwelling	Owners	Owners
cond25	Dwelling is a condominium	Owners	Owners
mortgn25	More than one mortgage	Owners	Mortgage payers
mortgm25	Monthly mortgage payment	Mortgage	Mortgage payers
prtxm25	Monthly property taxes	Property-tax and condo	Owners
condm25	Monthly condominium fee	Property-tax and condo	Condominium members
rnre25	Reduced rent and reason	Renters	Renters
rnpk25	Parking included in rent	Renters	Rent payers
rnht25	Heat included in rent	Renters	Rent payers
rnwa25	Water included in rent	Renters	Rent payers
rnec25	Electricity included in rent	Renters	Rent payers
rntv25	Cable TV included in rent	Renters	Rent payers
rnfg25	Refrigerator included in rent	Renters	Rent payers
rnst25	Stove included in rent	Renters	Rent payers
rnwd25	Washer and dryer included in rent	Renters	Rent payers
rnfu25	Furniture included in rent	Renters	Rent payers
rnno25	No amenities included in rent	Renters	Rent payers
rnbs25	Rent calculated on basis of income	Renters	Renters
rentm25	Monthly rent	Rent	Renters

N° 75F0002MIF au catalogue — N° 010
ISSN: 1707-2867
ISBN: 0-662-74645-7

Document de recherche

Série de documents de recherche - Revenu

Imputation générale des variables du logement (sans les services publics) dans l'Enquête sur la dynamique du travail et du revenu (EDTR)



par Georgina House

Division de la statistique du revenu
Division des méthodes d'enquêtes sociales
Immeuble Jean-Talon, Ottawa, K1A 0T6

Téléphone: 1 613 951-7355

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division de la statistique du revenu, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : (613) 951-7355; (888) 297-7355 : revenu@statcan.ca).

Pour obtenir des renseignements sur l'ensemble des données de Statistique Canada qui sont disponibles, veuillez composer l'un des numéros sans frais suivants. Vous pouvez également communiquer avec nous par courriel ou visiter notre site Web.

Service national de renseignements	1 800 263-1136
Service national d'appareils de télécommunications pour les malentendants	1 800 363-7629
Renseignements concernant le Programme des services de dépôt	1 800 700-1033
Télécopieur pour le Programme des services de dépôt	1 800 889-9734
Renseignements par courriel	infostats@statcan.ca
Site Web	www.statcan.ca

Renseignements pour accéder au produit

Le produit nº 75F0002MIF au catalogue est disponible gratuitement. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Nos produits et services.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois, et ce, dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1 800 263-1136. Les normes de service sont aussi publiées dans le site www.statcan.ca sous À propos de Statistique Canada > Offrir des services aux Canadiens.

Statistique Canada
Division de la statistique du revenu

Série de documents de recherche - Revenu

Imputation générale des variables du logement (sans les services publics) dans l'Enquête sur la dynamique du travail et du revenu (EDTR)

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2005

Tous droits réservés. Le contenu de la présente publication peut être reproduit, en tout ou en partie, et par quelque moyen que ce soit, sans autre permission de Statistique Canada sous réserve que la reproduction soit effectuée uniquement à des fins d'étude privée, de recherche, de critique, de compte rendu ou en vue d'en préparer un résumé destiné aux journaux, et/ou à des fins non commerciales. Statistique Canada doit être cité comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, il est interdit de reproduire quelque contenu de la présente publication, ou de l'emmagasiner dans un système de recouvrement, ou de le transmettre sous quelque forme et par quelque moyen que ce soit, reproduction électronique, mécanique, photographique, pour quelque fin que ce soit, sans l'autorisation écrite préalable des Services d'octroi de licences, Division du marketing Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Juillet 2005

N° 75F0002MIF au catalogue, vol. 10

ISSN: 1707-2867

ISBN: 0-662-74645-7

Périodicité : hors-série

Ottawa

This publication is available in English upon request (Catalogue no. 75F0002MIE).

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population, les entreprises, les administrations canadiennes et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques précises et actuelles.

Table des matières

Introduction 5
Imputation longitudinale 6
Imputation transversale par donneur 9
Module des types de logement et des modes d'occupation 10
Module des propriétaires et des locataires 10
Module des propriétaires 11
Module des prêts hypothécaires 11
Module des impôts fonciers et des frais de copropriété 12
Module des locataires 12
Module des loyers 12
Évaluation des méthodes d'imputation pour l'année de référence 2002 13
Graphique 1 : Histogramme du bassin des donneurs pour RENTM25 15
Graphique 2 : Histogramme du fichier après imputation (RENTM25<1500) pour RENTM25 15
Conclusion 17
Annexe 1 19
Liste des variables 19

Introduction

Depuis un certain temps, la Société canadienne d'hypothèques et de logement (SCHL[1]) utilise les données du Recensement de la population sur les caractéristiques du logement et les dépenses liées au logement. Bien que la source de données que constitue le recensement réponde dans une large mesure aux besoins de la SCHL, cet organisme du gouvernement fédéral a exploité les enquêtes-ménages annuelles de Statistique Canada pour obtenir des renseignements plus fréquents. Cela devait lui permettre d'avoir une image plus fidèle des tendances annuelles, et peut-être d'avoir un plus grand choix d'autres caractéristiques pour le recoupement des données sur le logement des ménages canadiens. En 2001, la SCHL a commencé à commanditer des questions supplémentaires à la fois dans l'Enquête sur la dynamique du travail et du revenu (EDTR) et dans l'Enquête sur les dépenses des ménages (EDM), à compter de l'année de référence 2002[2].

L'Enquête sur la dynamique du travail et du revenu (EDTR) est une enquête longitudinale qui date de 1993. Elle est destinée à mesurer les variations du bien-être économique des Canadiens ainsi que les facteurs touchant ces changements. La population cible est formée de toutes les personnes vivant au Canada, sauf les suivantes : les résidents du Yukon, des Territoires du Nord-Ouest et du Nunavut, les résidents des réserves, les personnes institutionnalisées et le personnel militaire en casernement.

L'échantillon de l'EDTR est formé de 2 panels. Chaque panel demeure dans l'enquête pour 6 années consécutives, et un nouveau panel y entre tous les 3 ans. En janvier suivant l'année de référence, les préposés aux interviews téléphoniques prennent contact avec les ménages de l'échantillon de l'EDTR. Ils recueillent des données démographiques pour chaque membre du ménage. Ils recueillent des données d'enquête complètes pour chaque personne admissible de plus de 15 ans qui se trouve dans le ménage. Ils posent des questions sur le travail (l'activité sur le marché du travail, l'expérience de travail, les épisodes de chômage et l'emploi même), le niveau de scolarité et les sources de revenu. À la fin de l'interview de janvier, ils disent aux répondants qu'on recommuniquera avec eux en mai pour obtenir des données sur leur revenu ainsi que sur certains postes de dépenses. Cependant, le répondant peut dès lors accorder à Statistique Canada la permission de récupérer toutes les données requises dans le fichier des données fiscales T1, ce qui élimine la nécessité d'une deuxième interview. La collecte des données sur le revenu est reportée en mai, de sorte que le répondant (venant alors juste de produire une déclaration de revenus) connaîtra mieux les données requises.

Bien que conçue à l'origine comme une enquête longitudinale, l'EDTR a toujours conservé la capacité de produire des estimations transversales. Toutes les personnes qui sont membres des ménages choisis pour l'EDTR au début de la première année de

1. La Société canadienne d'hypothèques et de logement (SCHL) est un organisme du gouvernement fédéral dont le mandat est de favoriser : la construction résidentielle, et la réparation et la modernisation des habitations existantes; l'accès à une diversité de logement abordables; l'amélioration des conditions de logement; la disponibilité de financement à moindre coût; et la prospérité du secteur de l'habitation.
2. Au moment de la rédaction du présent rapport, cette commandite se poursuit chaque année.

l'existence d'un panel font partie de l'échantillon longitudinal pour l'EDTR. À ce titre, ce sont ces particuliers qui font l'objet d'une observation longitudinale. Toute autre personne vivant dans un ménage avec une personne de l'échantillon longitudinal est appelée cohabitant. Les cohabitants vivant avec des personnes de l'échantillon longitudinal admissibles à l'échantillon transversal font aussi partie de l'échantillon transversal.

Pour plus de renseignements sur les concepts, les définitions et la conception de l'enquête, voir la publication de Statistique Canada intitulée « *Enquête sur la dynamique du travail et du revenu – Un aperçu de l'enquête* », http://www.statcan.ca :8096/bsolc/francais/bsolc?catno=75F0011X.

Pour répondre au besoin de données plus abondantes de la SCHL, Statistique Canada a ajouté plus de 20 questions sur le logement à l'interview de l'EDTR sur le travail. Certaines questions s'adressent aux propriétaires-occupants et aux locataires, certaines uniquement aux propriétaires-occupants, et d'autres uniquement aux locataires. L'EDTR a commencé à recueillir des renseignements sur les paiements hypothécaires, les impôts fonciers, les frais de copropriété, les loyers et ce qui est inclus dans le loyer.

À cause de la non-réponse à certaines questions particulières, il a fallu instituer l'imputation des variables du logement dans l'EDTR[3]. L'objet de l'imputation est de remplacer des données manquantes ou absentes par des valeurs qui donnent des estimations raisonnables. Deux méthodes d'imputation ont été utilisées : l'imputation longitudinale et l'imputation transversale par donneur. C'est ce que nous expliquons dans les deux sections suivantes.

Imputation longitudinale

L'imputation longitudinale tient compte, dans la mesure du possible, des renseignements du cycle précédent. L'efficacité de l'imputation est considérablement accrue lorsque les cycles consécutifs sont en étroite corrélation.

L'imputation longitudinale des variables du logement a été effectuée dans toute la mesure du possible. L'imputation longitudinale est préférable, surtout lorsque le ménage n'a pas déménagé. Nous supposons que, si le ménage n'a pas changé de résidence depuis l'année précédente, les renseignements pour la nouvelle année risquent d'être les mêmes que ceux de l'année précédente.

Puisque l'année de référence 2002 était l'année d'un nouveau panel, l'imputation longitudinale a été effectuée différemment, selon que le ménage faisait ou pas partie du premier cycle du nouveau panel. Pour les ménages qui ne faisaient pas partie du premier cycle du nouveau panel, les renseignements pour les trois variables du logement (mode d'occupation, type de logement et nombre de chambres à coucher) pouvaient s'obtenir des années précédentes. C'étaient les trois seules variables du logement qui n'étaient pas

3. L'EDTR a toujours fait l'imputation des valeurs manquantes pour les variables du revenu.

nouvelles dans l'EDTR pour l'année de référence 2002. (Les données sur le prêt hypothécaire sur le logement pouvaient aussi s'obtenir des années antérieures, mais il a été décidé de ne pas effectuer d'imputation longitudinale de cette variable, parce que la situation hypothécaire risque davantage de changer même s'il semble que l'adresse est restée la même.) On a imputé les renseignements manquants pour les ménages à l'aide des renseignements du même ménage pour les années précédentes si le code postal était toujours le même. C'était considéré comme une bonne indication que le ménage n'avait pas changé de logement.

L'Enquête sur la population active (EPA) et l'Enquête sur le loyer (un supplément de l'EPA) ont été exploitées comme source de renseignements pour l'imputation longitudinale pour les ménages faisant partie du premier cycle du nouveau panel[4]. On a imputé les renseignements manquants pour les ménages en utilisant les renseignements du même ménage selon les données de l'EPA si le code postal était le même. Pour utiliser l'Enquête sur le loyer, il a fallu vérifier si la variable du mode d'occupation dans l'EDTR indiquait que le ménage était toujours locataire.

L'imputation longitudinale a permis d'abaisser le nombre de ménages avec valeurs manquantes pour les variables qui étaient accessibles dans les années antérieures de l'EDTR. Pour d'autres, cependant, elle a eu une incidence faible ou nulle. Les nouvelles variables du logement pour lesquelles la seule source de renseignements était l'enquête sur le loyer ont donné très peu d'imputations longitudinales valides. C'est surtout parce que l'enquête sur le loyer n'était accessible que pour les ménages du cycle 4 qui étaient considérés comme locataires. Sur ce faible nombre, seuls ceux ayant des données valides dans l'Enquête sur le loyer et ceux dont le code postal concordait ont alors fait l'objet d'une imputation.

Pour l'année de référence 2003, l'imputation longitudinale a été effectuée sur toutes les variables du logement, pour les deux panels. On a imputé les renseignements manquants pour les ménages à l'aide des renseignements du même ménage de l'année de référence 2002 si le code postal était demeuré le même. Si le mode d'occupation était connu pour l'année de référence 2003, alors les renseignements manquants n'ont été imputés que si le mode d'occupation concordait également.

Le tableau 1 indique le pourcentage pondéré des valeurs manquantes pour chaque variable. Là où l'imputation longitudinale était possible, le pourcentage pondéré des valeurs manquantes après imputation longitudinale est également indiqué. Puisque l'imputation longitudinale n'était possible que pour certaines variables en 2002, certaines valeurs sont toujours manquantes. Certaines variables ne s'appliquent qu'à certains ménages (p. ex., propriétaires-occupants et locataires, propriétaires-occupants seulement, et locataires seulement); par conséquent, un grand nombre n'ont pas le même dénominateur. Ceux pour qui il faut imputer une valeur autre que « sans objet » constituent les différents dénominateurs. Pour les définitions des variables, voir l'annexe 1.

4. L'EDTR utilise les ménages qui sortent de l'EPA en décembre de l'année précédant l'année de référence de l'EDTR et en janvier de l'année de référence de l'EDTR.

Tableau 1 – Effet de l'imputation longitudinale pour 2002 et 2003

Variable	Groupe de ménages nécessitant l'imputation pour cette variable	Valeurs manquantes (%)				Nombre de ménages soumis à l'imputation longitudinale	
		Avant		Après			
		2002	2003	2002	2003	2002	2003
dwldet25	Propriétaires et locataires	6,9	7,4	0,6	0,5	1646	1871
dwtenr25	Propriétaires et locataires	6,6	7,4	0,4	0,5	1637	1908
dwltyp25	Propriétaires et locataires	6,9	7,4	0,6	0,5	1646	1871
rooms25	Propriétaires et locataires	6,9	8,0	2,7	0,6	1215	2031
opbu25	Propriétaires et locataires	8,1	9,4	8,0	0,8	9	2649
rnre25	Locataires	8,8	10,8	8,7	0,9	7	798
rnpk25	Payent un loyer	8,0	9,5	7,9	0,6	6	689
rnht25	Payent un loyer	8,0	9,5	7,9	0,6	4	689
rnwa25	Payent un loyer	8,0	9,5	7,9	0,6	5	689
rnec25	Payent un loyer	8,0	9,5	7,9	0,6	2	689
rntv25	Payent un loyer	8,0	9,5	7,9	0,6	1	689
rnfg25	Payent un loyer	8,0	9,5	7,9	0,6	5	689
rnst25	Payent un loyer	8,0	9,5	7,9	0,6	6	689
rnwd25	Payent un loyer	8,0	9,5	7,9	0,6	3	689
rnfu25	Payent un loyer	8,0	9,5	8,0	0,6	0	689
rnno25	Payent un loyer	8,0	9,5	8,0	0,6	0	689
rentm25	Locataires	12,1	14,4	10,2	2,6	109	946
repa25	Propriétaires et locataires	7,0	7,8	s.o.[2]	0,6	s.o.	1987
heat25	Propriétaires et locataires	11,6	12,8	s.o.	2,3	s.o.	2703
heatg25	Propriétaires et locataires	11,6	12,8	s.o.	2,3	s.o.	2703
opfm25	Propriétaires et locataires	3,4[1]	7,0	s.o.	0,7	s.o.	1716
mortg25	Propriétaires	7,7	8,1	s.o.	0,8	s.o.	1426
cond25	Propriétaires	6,5	7,1	s.o.	0,6	s.o.	1267
mortgn25	Paient une hypothèque	9,5	10,3	s.o.	1,9	s.o.	861
mortgm25	Paient une hypothèque	29,4	29,5	s.o.	8,6	s.o.	2297
prtxm25	Propriétaires	25,7	26,7	s.o.	5,4	s.o.	4546
condm25	Membres d'un condominium	13,8	15,7	s.o.	5,4	s.o.	95
rnbs25	Locataires	9,6	11,3	s.o.	1,2	s.o.	834

1. Dans l'année de référence 2002, seuls les ménages (de l'extérieur du Nouveau-Brunswick) qui ont un code postal indiquant une région rurale (deuxième chiffre du code postal = 0) se sont vu demander si le ménage exploitait une ferme dans sa propriété (OPFM25); tous les autres sont fixés à « sans objet ». Le vrai taux de non-réponse est de 16,6 %, avec dénominateur de 11 018.
2. Imputation longitudinale non disponible en 2002.

Imputation transversale par donneur

L'imputation par donneur consiste à définir un groupe de ménages partageant plusieurs caractéristiques avec le ménage soumis à l'imputation, puis à choisir l'un d'entre eux comme donneur. La valeur déclarée par le donneur remplace la zone manquante du ménage soumis à l'imputation. Ainsi, l'EDTR a imputé le revenu d'un particulier en choisissant le revenu d'une personne du même sexe et du même niveau d'instruction vivant dans la même province, faisant partie de la même tranche d'âge et ayant le même niveau d'instruction, le même type d'emploi (salarié, travailleur indépendant) et la même profession.

L'imputation par donneur a été effectuée pour les autres variables et pour tous les ménages où l'imputation longitudinale était impossible. L'imputation par donneur a été divisée en sept modules. Il s'agissait de faciliter l'utilisation des variables imputées dans un module ou comme variables auxiliaires d'appariement dans un autre. Les enregistrements imputés peuvent aussi servir de donneurs dans des cycles successifs d'imputation.

Chaque module a des listes différentes de variables auxiliaires et d'appariement. Les variables auxiliaires servent à créer des groupes d'imputation[5], alors qu'on utilise les variables d'appariement dans une fonction de pointage pour chercher le donneur le plus approprié dans un groupe d'imputation. Dans certains cas, les variables auxiliaires ont créé des groupes de receveurs[6] sans groupes de donneurs correspondants[7]; en l'occurrence, les variables auxiliaires ont été regroupées. Par exemple, dans tous les modules sauf un, la variable « province » a été regroupée en cinq régions. Lorsque le regroupement n'était pas suffisant pour créer des donneurs possibles pour tous les receveurs, on a transformé les variables auxiliaires en variables d'appariement. Dans chaque module, les variables figurant sur la liste des « variables à imputer » ont aussi servi de variables d'appariement lorsqu'une valeur pour cette variable était disponible. Par exemple, dans le module des types de logement et des modes d'occupation, si le type de logement était manquant et que le mode d'occupation ne l'était pas, on a tenté de trouver un donneur ayant une valeur concordante pour le mode d'occupation. Dans chaque groupe d'imputation, les donneurs étaient évalués par une fonction de pointage Cette fonction de pointage était basée sur la fonction de pointage utilisée dans l'imputation du revenu pour l'EDTR :

$$s(X,Y) = \sum_{k=1}^{K} p_k \, \mathrm{I}(X_k, Y_k), \text{ où } \mathrm{I}(X_k, Y_k) = \begin{cases} 1 & \text{si } X_k = Y_k \\ 0 & \text{autrement.} \end{cases}$$

5. Les donneurs et les receveurs (enregistrements nécessitant une imputation) sont séparés en groupes d'imputation. Chaque receveur au sein d'un groupe d'imputation est comparé à chaque donneur au sein du même groupe d'imputation.
6. Les groupes de receveurs sont les groupes, fondés sur les variables auxiliaires, de chaque module d'enregistrements nécessitant une imputation.
7. Les groupes de donneurs sont les groupes, fondés sur les variables auxiliaires, de chaque module d'enregistrements de donneurs possibles.

Noter que p_k est un poids nous permettant d'attribuer plus ou moins d'importance à la variable d'appariement k. Il a été décidé que toutes les variables d'appariement ont la même importance, d'où $p_k=1$. X_k est la valeur de la variable k du receveur et Y_k la valeur de la variable k du donneur.

On choisit alors un donneur en fonction de ce pointage. C'est le donneur ayant le plus haut pointage qui est choisi. S'il y a plusieurs donneurs ayant la valeur du pointage élevé, alors un donneur est choisi au hasard dans la liste. Pour les modules où il n'y a pas de variables d'appariement, un donneur est choisi au hasard dans l'ensemble du groupe d'imputation.

Les paragraphes suivants sont un résumé de chaque module d'imputation. Chaque tableau renferme une liste de variables à imputer avec les listes des variables auxiliaires et d'appariement. Il a été décidé que, pour les modules où il fallait imputer des variables numériques, celles qui se trouvaient dans le centile supérieur seraient exclues du bassin des donneurs. Cela se produit dans quatre modules différents; le module des propriétaires et des locataires, le module des prêts hypothécaires, le module des impôts fonciers et des frais de copropriété et, enfin, le module des loyers. C'était en partie pour diminuer le risque d'imputer une valeur supérieure au revenu du ménage pour des montants comme le loyer ou le paiement hypothécaire.

Module des types de logement et des modes d'occupation

Le module des types de logement et des modes d'occupation est le premier appliqué. De cette façon, chaque module appliqué utilise soit le type de logement soit le mode d'occupation comme variable auxiliaire ou d'appariement. Noter que le nombre de codes est indiqué entre crochets.

Enregistrements dans le module :	Tous les ménages
Variables à imputer :	Type de logement détaillé, type de logement, mode d'occupation
Variables auxiliaires :	Province (10), groupe de taille de région urbaine (3), nombre de personnes dans le ménage au 31 décembre (3)

Module des propriétaires et des locataires

Le module des propriétaires et des locataires vient ensuite. Les variables à imputer dans ce module sont les variables pour lesquelles tous les enregistrements exigent une valeur autre que « sans objet ». Dans le choix des donneurs possibles dans ce module, on a exclu tous ceux ayant une valeur de 9 ou plus pour le nombre de chambres à coucher.

Enregistrements dans le module :	Tous les ménages
Variables à imputer :	Nombre de chambres à coucher, réparations nécessaires au logement, combustible principal utilisé pour le

	chauffage, groupe de combustibles principaux utilisés pour le chauffage, le ménage exploite une ferme, le ménage exploite une entreprise
Variables auxiliaires :	Province (5), groupe de taille de région urbaine (3), groupe de types de logement (3), nombre de personnes dans le ménage au 31 décembre (3)
Variables d'appariement :	Propriété du logement (2)

Module des propriétaires

Vient ensuite le module des propriétaires. Ce module utilise les variables imputées des deux modules précédents à titre de variables auxiliaires. Tous les ménages qui ont une valeur de mode d'occupation indiquant qu'ils sont propriétaires du logement entrent dans ce module. Tous les autres enregistrements ayant des valeurs manquantes sont fixés à « sans objet ».

Enregistrements dans le module :	Propriétaires
Variables à imputer :	Prêt hypothécaire sur le logement, plus d'un prêt hypothécaire sur le logement, le logement est un logement en condominium
Variables auxiliaires :	Province (5), groupe de taille de région urbaine (3), groupe de types de logement (2), nombre de chambres à coucher (4)
Variables d'appariement :	Nombre de personnes au 31 décembre (3)

Module des prêts hypothécaires

Le module des prêts hypothécaires suit le module des propriétaires. C'est pour faire en sorte que seuls les ménages qui ont un prêt hypothécaire sur le logement (MORTG25) entrent dans ce module. Tous les autres pour lesquels les valeurs sont manquantes sont fixés à « sans objet ». Dans le choix des donneurs possibles dans ce module, tous ceux qui avaient une valeur pour un paiement hypothécaire mensuel courant supérieur à 3 000 $ ont été exclus.

Enregistrements dans le module :	Propriétaires ayant un prêt hypothécaire sur le logement
Variables à imputer :	Montant mensuel des paiements hypothécaires courants
Variables auxiliaires :	Province (5), type de logement (2), nombre de chambres à coucher (4), plus d'un prêt hypothécaire sur le logement (2)

Module des impôts fonciers et des frais de copropriété

Le module des impôts fonciers et des frais de copropriété peut être appliqué après le module de l'hypothèque. Tous les propriétaires-occupants entrent dans ce module. Tous les autres pour lesquels il y a des valeurs manquantes sont fixés à « sans objet ». Puisque la variable « logement en condominium » (COND25) est l'une des variables d'appariement des frais de copropriété, les propriétaires-occupants qui ne font pas partie d'un condominium verront leurs frais mensuels de copropriété remplacés, par imputation, par « sans objet ». En choisissant les donneurs possibles dans ce module, on a exclu tous ceux qui avaient un montant mensuel d'impôts fonciers supérieur à 500 $ ou une valeur de frais mensuels de copropriété supérieure à 800 $.

Enregistrements dans le module :	Propriétaire
Variables à imputer :	Montant mensuel des impôts fonciers, frais mensuels de copropriété
Variables auxiliaires :	Province (5), groupe de types de logement-propriétaires (2), logement en condominium (2)
Variables d'appariement :	Groupe de taille de région urbaine (3), nombre de chambres à coucher (4)

Module des locataires

Le module des locataires doit suivre le module des propriétaires et des locataires. Tous les ménages qui ont une valeur de mode d'occupation indiquant qu'ils sont locataires entrent dans ce module. Pour tous les autres pour lesquels des valeurs sont manquantes, on indique « sans objet ». Ceux qui sont locataires mais ne paient pas de loyer mensuel (RENTM25=0), ont des variables comme « l'électricité, le stationnement, etc., sont inclus dans le loyer » ont une valeur qui est fixée à « sans objet ».

Enregistrements dans le module :	Locataires
Variables à imputer :	… inclus dans le loyer (plusieurs éléments), la totalité ou une partie des meubles sont inclus, raison du loyer réduit, loyer calculé en fonction du revenu
Variables auxiliaires :	Province (5), groupe de taille de région urbaine (3), groupe de types de logement-locataires (3)
Variables d'appariement :	Nombre de chambres à coucher (4)

Module des loyers

Enfin, le dernier module d'imputation est celui des loyers. Tous les ménages qui payent un loyer mensuel entrent dans ce module. Tous les autres ménages pour lesquels des valeurs sont manquantes se voient imputer la valeur « sans objet ». En choisissant les

donneurs possibles dans ce module, on a exclu tous ceux qui avaient un montant mensuel de loyer supérieur à 1 500 $.

Enregistrements dans le module :	Locataires
Variables à imputer :	Montant mensuel du loyer
Variables auxiliaires :	Province (5), groupe de types de logement-locataires (3), nombre de chambres à coucher (4), raison du loyer réduit (2)
Variables d'appariement :	.....inclus dans le loyer (plusieurs éléments), nombre de chambres à coucher (4)

Il y a eu une vaste gamme de pourcentages de valeurs manquantes avant l'imputation par donneur pour l'année de référence 2002. Certaines variables ont un pourcentage relativement élevé de valeurs manquantes, de sorte que les bassins de donneurs sont restreints. Le tableau 1 indique le pourcentage pondéré des valeurs manquantes avant l'imputation transversale pour les années de référence 2002 et 2003.

Évaluation des méthodes d'imputation pour l'année de référence 2002

Pour évaluer le succès des méthodes d'imputation, il a fallu examiner les fréquences avant et après imputation pour les variables nominales et les moyennes avant et après imputation pour les variables continues. Les variations de moins de 1 % des pourcentages avant et après imputation et une variation de moins de 6 % des moyennes avant et après imputation ont été comptées comme résultat positif. On a aussi effectué une analyse de validation par recoupement à l'aide de 1 000 enregistrements de données complètes. Ces enregistrements ont été imputés. Puis, on a comparé les données initiales aux données imputées.

Après l'imputation à l'aide des méthodes décrites plus haut, plusieurs analyses ont été effectuées. Ces analyses avaient pour objet d'établir si les données obtenues étaient ou n'étaient pas semblables aux données d'origine. Les fréquences avant et après imputation ont été reprises sur toutes les variables catégoriques imputées. À titre d'exemple, le tableau 2 renferme des pourcentages avant et après imputation pour le type de logement. Avant l'imputation par donneur, 4 % des ménages étaient sans cette variable. Le tableau 2 indique que les pourcentages avant et après imputation sont très semblables. Toute différence qui existe est inférieure à 1 %, ce qui est très raisonnable.

Tableau 2 : Pourcentage avant et après imputation par donneur pour le type de logement

DWLDET25 Code fixé	Avant imputation par donneur (sans les valeurs manquantes dans le dénominateur)	Après imputation par donneur
01 Maison individuelle non attenante	66,5	65,9
02 Double	3,53	3,50
03 En rangée ou en terrasse	4,13	4,14
04 Duplex	4,02	4,02
05 Appartement de faible hauteur	14,0	14,4
06 Tour d'appartements	5,01	5,17
07 Hôtel, maison de chambres, camping	0,26	0,27
08 Maison mobile	2,42	2,42
09 Autres	0,15	0,17

Il est également important de regarder de près les variables pour lesquelles le pourcentage de valeurs manquantes est beaucoup plus élevé. Ainsi, le tableau 3 présente les pourcentages avant et après imputation pour la source principale de chauffage. Avant l'imputation par donneur, cette variable était manquante pour 24 % des ménages. Là encore, cependant, le tableau 3 indique que la différence des pourcentages est toujours inférieure à 1 %. Ces résultats sont typiques de ce que l'on a observé pour toutes les variables catégoriques.

Tableau 3 : Pourcentage avant et après imputation par donneur pour la source principale utilisée pour le chauffage

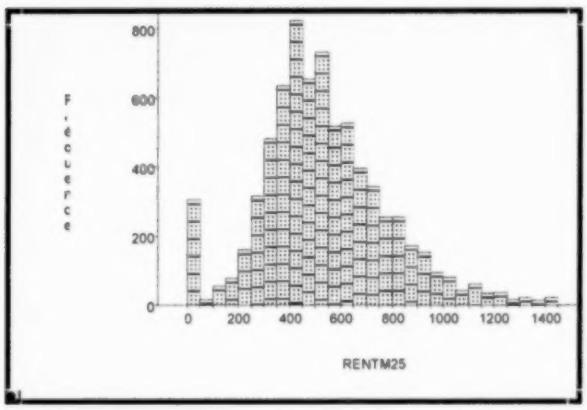
HEAT25 Code fixé	Pourcentage avant imputation par donneur (sans les valeurs manquantes dans le dénominateur)	Pourcentage après imputation par donneur
01 Mazout ou autre combustible liquide	17,1	16,7
02 Gaz canalisé (gaz naturel)	41,0	41,8
03 Gaz en bouteille (propane)	0,90	0,88
04 Électricité	33,0	33,5
05 Bois	7,39	7,07
06 Autres	0,06	0,06

Pour les variables continues, on a examiné les moyennes avant et après imputation. Cela a donné une bonne indication quant à la question de savoir si les données avaient ou non changé pour la peine. À titre d'exemple, voyons les moyennes avant et après imputation du montant du loyer mensuel (RENTM25) et des impôts fonciers (PRTXM25). La moyenne avant imputation de RENTM25 était de 604,12 $, et cette variable était manquante pour 29 % des ménages; après imputation, la moyenne était tombée

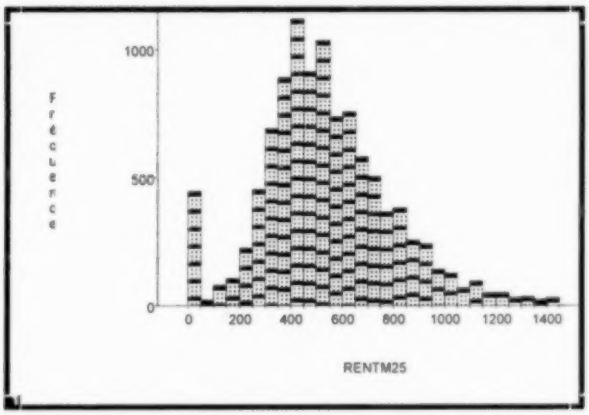
légèrement à 589,53 $. C'est une diminution de 2 %. La moyenne avant imputation PRTXM25 était de 160,68 $, et cette variable était manquante pour 35 % des ménages; après imputation, la moyenne a encore tombé légèrement, à 154,53 $. C'est une diminution de 4 %. Ces résultats sont caractéristiques de toutes les variables continues. Les réductions vont de 2 % à 6 %. Il faut prévoir une baisse de la moyenne, surtout parce que le centile supérieur du bassin de donneurs a été exclu.

On a aussi effectué un contrôle visuel. On a porté les variables continues sur un histogramme pour comparer le bassin des donneurs avec le fichier nouvellement imputé. Pour simplifier la comparaison, il a fallu exclure le centile supérieur des données de l'histogramme du fichier nouvellement imputé. Les histogrammes des graphiques 1 et 2 montrent que la distribution avant et après imputation est très semblable. On a ainsi comparé les histogrammes de toutes les variables continues, et l'exemple choisi est typique de ce que nous avons observé. C'est une bonne indication que les méthodes d'imputation choisies étaient bonnes, en ce qu'elles n'ont pas modifié les données pour la peine.

Graphique 1 : Histogramme du bassin des donneurs pour RENTM25



Graphique 2 : Histogramme du fichier après imputation (RENTM25<1500) pour RENTM25



À titre d'analyse complémentaire, on a effectué une validation par recoupement. Une validation par recoupement est une comparaison, appliquée aux ménages pour lesquels nous avons des données complètes, entre les valeurs imputées par les méthodes décrites plus haut et les valeurs déclarées. Une validation par recoupement est utile parce qu'elle simule le processus d'imputation d'une manière qui facilite la vérification de son exactitude. Un ensemble de 1 000 enregistrements pour lesquels il y avait des données complètes a été soumis à l'imputation par les méthodes décrites plus haut. Les données initiales de ces enregistrements ont été comparées aux données nouvellement imputées. Cette comparaison par validation par recoupement n'a été effectuée que pour les

variables catégoriques. Une valeur incorrecte a été imputée entre 10 % et 40 % du temps. La validation par recoupement a été effectuée plusieurs fois et les résultats qui suivent sont caractéristiques de ce qui a été observé.

Lorsqu'on considère les résultats de la validation par recoupement, il faut retenir que l'imputation s'effectue en modules séquentiels. Le module des types de logement et des modes d'occupation est appliqué d'abord, si bien que les 11 % de ménages dont le mode d'occupation est incorrect seront alors incorrects pour la plupart des autres modules. Ainsi, ceux qui sont considérés comme propriétaires sont les seuls pour lesquels serait imputée une valeur pour le prêt hypothécaire sur le logement (MORTG25); tous les autres se verraient imputer la valeur « sans objet ».

Pour fins de comparaison, on a effectué des imputations aléatoires pour les 1 000 mêmes ménages. On a imputé une valeur aléatoire à chaque variable, en fonction des fréquences initiales.

Puisque l'imputation est séquentielle, l'imputation d'un module s'effectue avant l'imputation d'un autre module. Cela génère deux façons de procéder avec l'analyse. L'imputation aurait pu utiliser les valeurs créées pendant la validation par recoupement dans le fichier précédent, ou aurait pu utiliser les nouvelles valeurs qui ont été attribuées aléatoirement dans le fichier précédent. Pour maintenir une analyse logique des variables imputées, nous utilisons les données de la validation par recoupement. De cette façon, les répondants qui étaient au départ des locataires sont demeurés locataires, si bien que les variables applicables aux locataires, comme « chauffage inclus dans le loyer » (RNHT25) auraient une valeur imputée valide plutôt qu'une valeur « sans objet », comme ce serait le cas si la valeur imputée aléatoirement dans un module antérieur en faisait des propriétaires. Par conséquent, ceux qui sont considérés comme incorrectement imputés par la méthode aléatoire sont fondés sur le même sous-ensemble de la population que ceux de la validation par recoupement.

Il ressort de ces résultats qu'il y a une différence entre les variables pour lesquelles l'imputation longitudinale a donné un résultat positif et les variables pour lesquelles il n'y a pas eu de résultat positif. Par exemple, l'imputation aléatoire a été effectuée sur DWTENR25. Elle l'a été plusieurs fois et, typiquement, les données ont été imputées incorrectement 40 % du temps, ce qui est beaucoup plus qu'avec les méthodes d'imputation exposées ici. À titre d'exemple d'imputation aléatoire d'une variable où l'imputation longitudinale n'était pas disponible, on peut considérer la variable MORTG25. Les résultats révèlent que la méthode d'imputation ne donne qu'une légère amélioration par rapport à la méthode d'imputation aléatoire. Typiquement, la valeur incorrecte a été imputée 43 % du temps. Il semblerait, au vu de ces résultats, que la méthode d'imputation longitudinale est bien meilleure que l'imputation aléatoire, mais que la méthode d'imputation transversale par donneur n'est que légèrement meilleure. Le tableau 4 est un tableau de la différence en pourcentage entre les données initiales et les données imputées pour la validation par recoupement, avec les résultats typiques des imputations aléatoires.

Tableau 4 : Pourcentage de catégories incorrectement imputées pour deux stratégies d'imputation

Variable	Imputation transversale (%)	Imputation aléatoire (%)
Dwtenr25	11	40
Dwldet25	20	54
Mortg25	40	43
Repa25	35	36
Heat25	39	67
Heatg25	39	67
Obpu25	16	16
Mortgn25	38	38
Cond25	13	17
Rnpk25	17	18
Rnwa25	16	19
Rnec25	19	21
Rnht25	20	22
Rntv25	15	15
Rnfg25	16	20
Rnst25	16	21
Rnwd25	16	17
Rnfu25	13	13
Rnno25	12	12
Rnre25	17	18
Rnbs25	17	19

Conclusion

Étant donné la non-réponse initialement élevée pour les variables du logement, il a été utile de mettre en place une méthodologie d'imputation. L'un des points forts de la procédure d'imputation décrite est que le résultat final est un ensemble de données sans valeurs manquantes et avec convergence interne des enregistrements imputés.

Il ressort de ces analyses que les méthodes d'imputation choisies ont donné d'excellents résultats. Les fréquences avant et après imputation pour les variables catégoriques et les moyennes avant et après imputation pour les variables continues ne font ressortir que des variations relativement faibles de moins de 1 % et 6 %, respectivement. Comme confirmation supplémentaire, les histogrammes avant et après imputation se ressemblent beaucoup. Cela confirme que la procédure d'imputation a bien fonctionné pour préserver les distributions à une variable.

Une question plus complexe est celle de savoir dans quelle mesure les distributions à une variable se sont trouvées modifiées. Pour répondre à cette question, nous avons examiné

les estimations pour certains groupes déterminés. Ainsi, puisque l'âge n'est pas explicitement pris en compte dans l'imputation, on risque de perdre la corrélation existante entre l'âge et le mode d'occupation. Selon des analyses effectuées sur les données sur l'âge du membre du ménage le plus âgé, le processus d'imputation ne change rien aux estimations de propriétaires-occupants selon l'âge. La distribution de l'âge du plus vieux membre du ménage chez les propriétaires-occupants est très semblable avant et après imputation. L'âge moyen de la personne ayant le plus haut revenu net dans les ménages ayant un paiement hypothécaire mensuel imputé par opposition à non imputé est très semblable. C'est une preuve de plus que la méthode d'imputation est très bonne. Puisqu'il est impossible de vérifier toutes les variables pour lesquelles il pourrait y avoir une corrélation, on ne peut être sûr qu'aucune des estimations pour les groupes particuliers n'est touchée par ce qui semble être, selon une analyse de validation par recoupement, un taux d'erreur élevé.

En somme, bien que la mesure du taux d'erreur pour certaines variables soit élevée à cause de la faiblesse des modèles d'imputation sous-jacents, l'incidence sur les distributions à une variable et certaines distributions à variables multiples semble plutôt limitée. Autrement dit, bien que les taux de non-réponse aient été au départ élevés, on a pu remplacer les valeurs manquantes sans causer de changements importants à l'ensemble des estimations.

Annexe 1

Liste des variables

Variable	Description de la variable	Module	Groupe de ménages pour lesquels cette variable doit être imputée
dwldet25	Type de logement détaillé	Type de logement/ mode d'occupation	Propriétaires et locataires
dwtenr25	Propriété du logement	Type de logement/ mode d'occupation	Propriétaires et locataires
dwltyp25	Type de logement	Type de logement/ mode d'occupation	Propriétaires et locataires
rooms25	Nombre de chambres à coucher	Propriétaires/locataires	Propriétaires et locataires
Opbu25	Entreprise dans la propriété	Propriétaires/locataires	Propriétaires et locataires
Repa25	Réparations nécessaires	Propriétaires/locataires	Propriétaires et locataires
Heat25	Combustible principal utilisé pour le chauffage	Propriétaires/locataires	Propriétaires et locataires
heatg25	Groupe de combustibles principaux utilisés pour le chauffage	Propriétaires/locataires	Propriétaires et locataires
Opfm25	Ferme dans la propriété	Propriétaires/locataires	Propriétaires et locataires
mortg25	Prêt hypothécaire sur le logement	Propriétaires	Propriétaires
Cond25	Logement en condominium	Propriétaires	Propriétaires
mortgn25	Plus d'un prêt hypothécaire	Propriétaires	Débiteurs hypothécaires
mortgm25	Paiement hypothécaire mensuel	Prêt hypothécaire	Débiteurs hypothécaires
prtxm25	Impôts fonciers mensuels	Impôts fonciers et frais de copropriété	Propriétaires
condm25	Frais mensuels de copropriété	Impôts fonciers et frais de copropriété	Membres d'un condominium
Rnre25	Loyer réduit et raison	Locataires	Locataires
Rnpk25	Stationnement inclus dans le loyer	Locataires	Payent un loyer
Rnht25	Chauffage inclus dans le loyer	Locataires	Payent un loyer
rnwa25	Eau incluse dans le loyer	Locataires	Payent un loyer
Rnec25	Électricité incluse dans le loyer	Locataires	Payent un loyer
Rntv25	Câble inclus dans le loyer	Locataires	Payent un loyer
Rnfg25	Réfrigérateur inclus dans le loyer	Locataires	Payent un loyer

Rnst25	Cuisinière incluse dans le loyer	Locataires	Payent un loyer
rnwd25	Laveuse et sécheuse incluses dans le loyer	Locataires	Payent un loyer
Rnfu25	Meubles inclus dans le loyer	Locataires	Payent un loyer
Rnno25	Pas de commodités incluses dans le loyer	Locataires	Payent un loyer
Rnbs25	Loyer calculé en fonction du revenu	Locataires	Locataires
rentm25	Loyer mensuel	Loyer	Locataires